Chapter 12

Comparing bilingual word embeddings to translation dictionaries for extracting multilingual collocation equivalents

Marcos Garcia
Universidade da Coruña

This chapter introduces a strategy for the automatic extraction of multilingual collocation equivalents which takes advantage of parallel corpora to train bilingual word embeddings. First, monolingual collocation candidates are retrieved using syntactic dependencies and standard association measures. Then, the distributional models are applied to search for equivalents of the elements of each collocation in the target languages. The proposed method extracts not only collocation equivalents with direct translations between languages, but also other cases where the collocations in the two languages are not literal translations of each other. Several experiments – evaluating collocations with five syntactic patterns – in English, Spanish, and Portuguese show that this approach can effectively extract large sets of bilingual equivalents with an average precision of about 85%. Moreover, preliminary results on comparable corpora suggest that the distributional models can be applied for identifying new bilingual collocations in different domains. This strategy is compared to both hand-crafted bilingual dictionaries and to probabilistic translation dictionaries learned from the same resources as the bilingual word embeddings, showing that it achieves much larger recall values while keeping high precision results.

1 Introduction

MWEs have been repeatedly classified as an important problem for developing Natural Language Processing (NLP) tools, as well as to automatically analyze

Marcos Garcia. 2018. Comparing bilingual word embeddings to translation dictionaries for extracting multilingual collocation equivalents. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 319–342. Berlin: Language Science Press. DOI:10.5281/zenodo.1469571

linguistic utterances (Sag et al. 2002). Among the different types of MWEs, processing collocations in an automatic way may pose various problems due to their intrinsic properties such as compositionality or unpredictability (Mel'čuk 1998).

From a theoretical perspective, there are at least two main views on collocations. On the one hand, there is a tendency to consider any frequent pair of words to be a collocation (Smadja 1993; Evert & Kermes 2003; Kilgarriff 2006). On the other hand, the phraseological tradition needs both a lexical restriction and a syntactic relation to consider two lexical units as a collocation.[1] From this phraseological point of view, a collocation is a restricted binary co-occurrence of lexical units between which a syntactic relation holds, and that one of the lexical units (the BASE) is chosen according to its meaning as an isolated lexical unit, while the other (the COLLOCATE) is selected depending on the base and the intended meaning of the co-occurrence as a whole, rather than on its meaning as an isolated lexical unit (Mel'čuk 1998). Thus, a noun in English such as *picture* (as a direct object) requires the verb *to take* (and not *to do*, or *to make*) in the phrase *take a picture*, while *statement* selects *to make* (*make a statement*).

In a bilingual (or multilingual) scenario, equivalent collocations are needed to produce more natural utterances in the target language(s). In this regard, the referred noun *fotografia* 'picture' would select the verb *tirar* 'to remove' in Portuguese (*tirar uma fotografia*). Similarly the Spanish *vino* 'wine' would require the adjective *tinto* (*vino tinto*), which is not the main translation of *red* (*red wine*).

The unpredictability of these structures poses problems for tasks such as machine translation, whose performance can benefit from lists of multilingual collocations or transfer rules for these units (Orliac & Dillinger 2003). In areas like second language learning, it has been shown that even advanced learners need to know which word combinations are allowed in a specific linguistic variety (Altenberg & Granger 2001; Alonso-Ramos et al. 2010). Thus, obtaining resources of multilingual equivalent collocations could be useful for a variety of applications such as those mentioned above. However, this kind of resource is scarce, and constructing them manually requires a large effort from expert lexicographers.

Since the 1990s, a number of approaches were implemented aimed at extracting bilingual collocations, both from parallel corpora (Kupiec 1993; Smadja et al. 1996; Wu & Chang 2003), and from comparable or even from non-related monolingual resources (Lü & Zhou 2004; Rivera et al. 2013), often combining statistical approaches with the use of bilingual dictionaries to find equivalents of each base.

[1] An overview of different views on collocations – both from theoretical and practical perspectives – can be found in Seretan (2011).

This chapter explores the use of distributional semantics (by means of bilingual word embeddings) for identifying bilingual equivalents of monolingual collocations: On the one hand, monolingual collocation candidates are extracted using a harmonized syntactic annotation provided by Universal Dependencies (UD),[2] as well as standard measures for lexical association. On the other hand, bilingual word embeddings are trained using lemmatized versions of noisy parallel corpora. Finally, these bilingual models are employed to search for semantic equivalents of both the base and the collocate of each collocation.

Several experiments using the OpenSubtitles2016 parallel corpora (Lison & Tiedemann 2016) in English, Portuguese, and Spanish show that the proposed method successfully identifies bilingual collocation equivalents with different patterns: *adjective-noun*, *noun-noun*, *verb-object*, *verb-subject*, and *verb-adverb*. Furthermore, preliminary results in comparable corpora suggest that the same strategy can be applied in this kind of resources to extract new pairs of bilingual collocations. In this regard, this chapter is an extended version of a previous work on bilingual collocation extraction (Garcia et al. 2017), including new collocation patterns and a larger evaluation which compares the proposed approach to probabilistic translation dictionaries (Hiemstra 1998; Simões & Almeida 2003).

Apart from this introduction, §2 includes a review of previous work on collocation extraction, especially on papers dealing with bilingual resources. Then, §3 and §4 present and evaluate the method, respectively. Finally, some conclusions and further work are discussed in §5.

2 Previous studies on collocation extraction

The extraction of monolingual collocation candidates (as well as other MWEs) from corpora is a well-known topic in corpus and computational linguistics and was the focus of a significant body of work in different languages.

In this respect, most strategies use statistical association measures on windows of n-grams with different sizes (Church & Hanks 1990; Smadja 1993). Other methods, such as the one presented by Lin (1999), started to apply dependency parsing to better identify combinations of words which occur in actual syntactic relations.

More recently, the availability of better parsers allowed researchers to combine automatically obtained syntactic information with statistical methods to extract collocations more accurately (Evert 2008; Seretan 2011).

[2] http://universaldependencies.org/

A different perspective on collocation extraction focuses not only on their retrieval, but on semantically classifying the obtained collocations, in order to make them more useful for NLP applications (Wanner et al. 2006; 2016).

Concerning the extraction of bilingual collocations, most works rely on parallel corpora to find the equivalent of a collocation in a target language. In this regard, Smadja (1992) and Smadja et al. (1996) first identify monolingual collocations in English (the source language), and then use Mutual Information (mi) and the Dice coefficient to find the French equivalents of the source collocations.

Kupiec (1993) also uses parallel corpora to find noun phrase equivalents between English and French. Their method consists of applying an expectation maximization (EM) algorithm to previously extracted monolingual collocations. Similarly, Haruno et al. (1996) obtain Japanese-English chunk equivalents by computing their mi scores and taking into account their frequency and position in the aligned corpora.

Another work which uses parallel corpora is presented by Wu & Chang (2003). The authors extract Chinese and English n-grams from aligned sentences by computing their log-likelihood ratio. Then, the competitive linking algorithm is used to decide whether each bilingual pair actually corresponds to a translation equivalent.

Seretan & Wehrli (2007) took advantage of syntactic parsing to extract bilingual collocations from parallel corpora. The strategy consists of first extracting monolingual collocations using log-likelihood, and then searching for equivalents of each base using bilingual dictionaries. The method also uses the position of the collocation in the corpus, and relies on the syntactic analysis by assuming that equivalent collocations will occur with the same syntactic relations within the collocations in both languages.

Rivera et al. (2013) present a framework for bilingual collocation retrieval that can be applied (using different modules) to both parallel and comparable corpora. As in other works, monolingual collocations based on n-grams are extracted in a first step, and then bilingual dictionaries (or WordNet, in the comparable corpora scenario) are used to find the equivalents of the base in the aligned sentence or in a small window of adjacent sentences of the source collocation.

A different approach, which uses non-related monolingual corpora for finding bilingual collocations, was presented in Lü & Zhou (2004). Here, the authors apply dependency parsing and the log-likelihood ratio for obtaining English and Chinese collocations. Then, they search for translations using word translation equivalents with the same dependency relation in the target language (using the EM algorithm and a bilingual dictionary).

Although not focused on collocations, Fung (1998) applied methods based on distributional semantics to build bilingual lexica from comparable corpora. This approach takes into account that in this type of resources the position and the frequency of the source and target words are not comparable, and also that the translations of the source words might not exist in the target document.

Similarly, the strategy presented in this chapter leverages noisy parallel corpora for building bilingual word embeddings. However, with a view to applying it to other resources such as comparable corpora, it identifies equivalents without using information about the position of the collocations or their comparative frequency in the corpora. Furthermore, it does not take advantage of external resources such as bilingual dictionaries, making it easy to extend to other languages. Garcia et al. (2018) had introduced a naive version of this approach, including experiments in Portuguese and Spanish with just one collocation pattern.

3 A new method for bilingual collocation extraction

This section presents the proposed method for automatically extracting bilingual collocations from corpora. First, the approach for identifying candidates of monolingual collocations using syntactic dependencies is briefly described. Then, the process of creating the bilingual word embeddings is shown, followed by the strategy for discovering the collocation equivalents between languages.

3.1 Monolingual dependency-based collocation extraction

Early works on n-gram based collocation extraction already pointed out the need for syntactic analysis to better identify collocations in corpora (Smadja 1993; Lin 1999). Syntactic analysis can, on the one hand, avoid the extraction of syntactically unrelated words which occur in small context windows. On the other hand, it can effectively establish a relation between lexical items occurring in long-distance dependencies (Evert 2008).

Besides, the method presented in this chapter assumes that most bilingual equivalents of collocations bear the same syntactic relation in both the source and the target languages, although it is not always the case (Lü & Zhou 2004).

In order to better capture the syntactic relations between the base and the collocate of each collocation, the strategy uses state-of-the-art dependency parsing. Apart from that, and aimed at obtaining harmonized syntactic information between languages, the method relies on Universal Dependencies annotation,

which makes it possible to use the same strategy for extracting and analyzing the collocations in multiple languages.

3.1.1 Preprocessing:

Before extracting the collocation candidates from each corpus, a pipeline of NLP tools is applied in order to annotate the text with the desired information. Thus, the output of this process consists of a parsed corpus in CoNLL-U format,[3] where each word is assigned to its surface form, its lemma, its POS-tag and morphosyntactic features, its syntactic head as well as the UD relation of the word in context.

From this analyzed corpus, the word pairs belonging to the desired relations (collocation candidates) are extracted. We keep their surface forms, POS-tags, and other syntactic dependents which may be useful for the identification of potential collocations. Besides, a list of triples is retained in order to apply association measures, containing (i) the syntactic relation, (ii) the head, and (iii) the dependent (using their lemmas together with the POS-tags). Thus, from a sentence such as *John took a great responsibility*, the following triples (among others) are obtained:

NSUBJ(take$_{\text{VERB}}$,John$_{\text{PROPN}}$)
AMOD(responsibility$_{\text{NOUN}}$,great$_{\text{ADJ}}$)
DOBJ(take$_{\text{VERB}}$,responsibility$_{\text{NOUN}}$)

This information, along with the corpus size and the frequency of the different elements of the potential collocations, is stored in order to rank the candidates.

3.1.2 Collocation patterns:

In this chapter, candidates of five different syntactic patterns of collocations are extracted in three languages, Spanish (ES), Portuguese (PT), and English (EN):[4]

- Adjective—Noun (AMOD): these candidates are pairs of adjectives (as collocates) and nouns (as bases) where the former syntactically depends of the latter in a AMOD relation. Example: **killer**$_{\text{base}}$;**serial**$_{\text{collocate}}$.
- Noun—Noun (NMOD): this pattern consists of two common nouns related by the NMOD relation, where the head is the base and the dependent is

[3] http://universaldependencies.org/format.html

[4] In this chapter we address the European variety of Portuguese. However, even if we use a European Portuguese corpus (see §4), it contains some texts in the Brazilian dialect.

the collocate (optionally with a CASE marking dependent preposition: *of* in English, *de* in Portuguese and Spanish). Example: **rage**$_{\text{base}}$;**fit**$_{\text{collocate}}$.[5]

- Verb—Object (VOBJ): *verb-object* collocations consist of a verb (the collocate) and a common noun (the base) occurring in a DOBJ relation. Example: **care**$_{\text{base}}$;**take**$_{\text{collocate}}$.
- Subj—Verb (VSUBJ): the VSUBJ collocation pattern contains a common noun (the base, acting as a subject) and the verb it depends on (the collocate). Example: **ship**$_{\text{base}}$;**sink**$_{\text{collocate}}$.
- Verb—Adverb (ADVMOD): in this case, a collocate adverb modifies a verb (the base) in an ADVMOD relation. Example: **want**$_{\text{base}}$;**really**$_{\text{collocate}}$.

3.1.3 Identification of candidates:

For each of the five patterns of collocations, a list of potential candidates for the three languages is extracted. After that, the candidates are ranked using standard association measures that have been widely used in collocation extraction (Evert 2008).

In the current experiments, two statistical measures were selected, whose results complement each other: T-SCORE, which prefers frequent dependency pairs, and has been proved useful for collocation extraction (Krenn & Evert 2001), and MUTUAL INFORMATION, which is useful for a large corpus, even if it tends to assign high scores to candidates with very low-frequency (Pecina 2010).

The output of both association measures is merged in a final list for each language and collocation pattern, defining thresholds of $t\text{-}score \geq 2$ and $MI \geq 3$ (Stubbs 1995), and extracting only collocations with a frequency of $f \geq 10$. This large value was defined to reduce the extraction of incorrect entries from a noisy corpus and from potential errors of the automatic analysis.

It must be noted that, since these lists of monolingual collocations have been built based on statistical measures of collocability, their members need not be *bona fide* collocations in the phraseological meaning. Thus, the lists can include idioms, e.g., *kick the bucket*, quasi-idioms, e.g., *big deal*, (Mel'čuk 1998), or free combinations, e.g., *buy a drink*.

[5]Some collocations belonging to this pattern are analyzed in UD – mainly in English – using the COMPOUND relation. These are not extracted in the experiments performed in this chapter.

3.2 Bilingual word embeddings

Word embeddings are low-dimensional vector representations of words which capture their distributional context in corpora. Even though distributional semantics methods have been largely used in previous years, approaches based on word embeddings gained popularity with the publication of *word2vec* (Mikolov et al. 2013). Based on the *Skip-gram* model of *word2vec*, Luong et al. (2015) proposed *BiSkip*, a model of word embeddings which learns bilingual representations using aligned corpora, thus being able to predict words crosslinguistically.

The method presented in this chapter uses lemmas instead of surface forms to identify the collocation candidates, so the bilingual models of word embeddings are also trained on lemmatized corpora. Therefore, the raw parallel corpus is lemmatized keeping the original sentence alignment.

The bilingual models are built using *MultiVec*, an implementation of *word2vec* and *BiSkip* (Berard et al. 2016). As the approach is evaluated in three languages, three different bilingual models are needed: Spanish-English, Portuguese-English, and Spanish-Portuguese.

As it will be shown, the obtained models can predict the similarity between words in bilingual scenarios by computing the cosine similarity between their vectors. As the models learn the distribution of single words (lemmas), they deal with different semantic phenomena such as polysemy or homonymy. Concerning collocations, this means that, ideally, the bilingual models could predict not only the equivalents of a base, but also to capture the (less close) semantic relation between the bilingual collocates, if they occur frequently enough in the data.

3.3 Bilingual collocation alignment

In order to identify the bilingual equivalent of a collocation in a target language, the method needs (i) lists of monolingual collocations (ideally obtained from similar resources), and (ii) a bilingual *source-target* model of word embeddings.

With these resources, the following strategy is applied: For each collocation in the source language (e.g., *lío*$_{base}$ *tremendo*$_{collocate}$ 'huge mess' in Spanish) the system selects its base and obtains – using the bilingual model – the *n* most similar lemmas in the target language (where *n=5* in the experiments performed in this chapter): *trouble*, *mess*, etc. Then, starting from the most similar lemma, we search in the target list for collocations containing the equivalents of the base (*trouble*$_{\text{base}}$ *little*$_{\text{collocate}}$, *trouble*$_{\text{base}}$ *deep*$_{\text{collocate}}$, *mess*$_{\text{base}}$ *huge*$_{\text{collocate}}$, *mess*$_{\text{base}}$ *fine*$_{\text{collocate}}$, etc.). If a collocation with a base equivalent is found, the cosine similarity between both collocates (*tremendo* versus *little*, *deep*, *huge*, and *fine*)

is computed, and they are selected as potential candidates if their similarity is higher than a given threshold (empirically defined in this chapter as 0.65), and if the target candidate is among the n most similar words of the source collocate (again, n=5). Finally, if these conditions are met, the source and target collocations are aligned, assigning the average distance between the bases and the collocates as a confidence value, as in the following Spanish-English example: $lío_{base}$ $tremendo_{collocate}$ = $mess_{base}$ $huge_{collocate}$ $\longrightarrow$ 0.721.

4 Evaluation

This section presents the experiments carried out in order to evaluate the proposed distributional method (henceforth DIS) in the three analyzed languages, using the five collocation patterns defined in §3.1. The approach presented in this chapter is compared to a baseline system (BAS), which uses hand-crafted bilingual dictionaries, and to probabilistic translation dictionaries (NAT).[6]

Corpora: Monolingual collocations were extracted from a subset of the OpenSubtitles2016 corpus (Lison & Tiedemann 2016), which contains parallel corpora from TV and Movie subtitles. This resource was selected because it is a large and multilingual parallel corpus likely to contain different types of collocations, also from an informal register, thus being useful for comparative studies.[7]

From the English, Spanish, and Portuguese corpora, those senteces which appear in the three languages were selected, for a total of 13,017,016 sentences. These sentences were tokenized, lemmatized and POS-tagged with a multilingual NLP pipeline (Garcia & Gamallo 2015), obtaining three corpora of about 91M (ES and PT), and about 98M (EN) tokens. The resulting data were enriched with syntactic annotation using statistical models trained with MaltParser (Nivre et al. 2007) on version 1.4 of the UD treebanks (Nivre et al. 2016).

Collocations: From each corpus, five patterns of collocation candidates were extracted: AMOD, NMOD, VOBJ, VSUBJ, and ADVMOD. For each language and pattern, a single list of collocations was obtained by merging the MI and T-SCORE outputs as explained in §3.1. Table 1 shows the number of filtered collocations in each case (*colls*).

[6] The extractions of these three methods are available at http://www.grupolys.org/~marcos/pub/pmwe-dis.tar.bz2

[7] Note, however, that OpenSubtitles2016 includes non-professional translations with some noisy elements such as typos or case inconsistencies, among others.

Table 1: Number of unique input dependencies for each syntactic pattern (*deps*), and final monolingual collocation candidates (*colls*).

Lg	AMOD		NMOD		VOBJ		VSUBJ		ADVMOD	
	deps	*colls*	*deps*	*colls*	*deps*	*colls*	*deps*	*colls*	*deps*	*colls*
ES	373K	13,870	644K	5,673	423K	17,723	287K	4,914	124K	5,526
PT	361K	12,967	709K	5,643	544K	20,984	283K	3,927	142K	6,660
EN	381K	14,175	517K	3,133	483K	15,492	264K	2,663	162K	6,711

Another version of each corpus was created only with the lemma of each token, keeping the original sentence alignments. These corpora were used for training three bilingual word embeddings with MultiVec, with 100 dimensions and a window-size of eight words: ES-EN, ES-PT, and PT-EN.[8]

Baseline (BAS): The performance of the method described in §3.3 was compared to a baseline which follows the same strategy, but uses bilingual dictionaries instead of the word embeddings models. Thus, the BAS method obtains the equivalents of both the base and the collocate of a source collocation, and verifies whether there is a target collocation with the translations. The bilingual dictionaries provided by the *apertium* project were used for these experiments (Forcada et al. 2011).[9]

The Spanish-Portuguese dictionary has 14,364 entries, and the Spanish-English one contains 34,994. The Portuguese-English dictionary (not provided by *apertium*) was automatically obtained by transitivity from the two other lexica, with a size of 9,160 pairs.

Probabilistic translation dictionaries (NAT): The distributional method was also compared to probabilistic translation dictionaries. Probabilistic dictionaries are bilingual resources which contain, for each word in a source language, possible translations in the target language together with the probability of the translation being correct. To obtain these dictionaries NATools was used, which is a set of tools to work with parallel corpora that can be utilized for different tasks such as sentence and word alignment, or to extract bilingual translation dictionaries by means of statistical methods (Simões & Almeida 2003). The probabilistic dictionaries are obtained by applying the EM algorithm on sparse matrices of

[8]These models are available at http://www.grupolys.org/~marcos/pub/mwe17_models.tar.bz2

[9]SVN revision 75,477, https://svn.code.sf.net/p/apertium/svn/

bilingual word co-occurrences, previously built from parallel corpora (Hiemstra 1998).

For a better comparison to the DIS model, NAT dictionaries were extracted from the same lemmatized resources used for training the bilingual word embeddings. Thus, this method only differs from DIS in the bilingual resources used to search for equivalents of the bases and the collocates.[10]

4.1 Results

With a view to knowing the performance of BAS, NAT, and DIS in the different scenarios, 100 bilingual collocation pairs were randomly selected from each language and pattern, creating a total of 45 lists (15 from each of the three methods).[11]

Three reviewers worked during the evaluation process. Each bilingual collocation pair was labeled as (i) correct, (ii) incorrect, or (iii) dubious, which includes pairs where the translation might be correct in some contexts even if they were not considered faithful translations.[12] Correct collocation equivalents are those pairs where the monolingual extractions were considered correct, both in terms of co-occurrence frequency and of collocation pattern classification, and whose translations were judged by the reviewers as potential translations in a real scenario. Two reviewers labeled each collocation pair in the BAS and DIS outputs, achieving 92% and 83% inter-annotator agreement, respectively, with an average $\kappa = 0.39$, which indicates the difficulty of this kind of annotation. Pairs with correct/incorrect disagreement were discarded for the evaluation. Those with at least one dubious label were checked by a third annotator, deciding in each case whether they were correct, incorrect, or dubious. This third annotator evaluated the outputs of NAT using exactly the same guidelines.

From these data, the precision values for each case were obtained by dividing the number of correct collocation equivalents by the number of correct, incorrect, and dubious cases (so dubious cases were considered incorrect). Recall (r) was obtained by multiplying the precision values (p) for the number of extracted equivalents (e), and dividing the result by the lowest number of input collocations for each pair (i, see Table 1). For instance, the Spanish-Portuguese baseline

[10] After preliminary evaluations, the translation probability thresholds of both lexical units were empirically defined as 0.1.

[11] Except for baseline extractions with less than 100 elements, where all of them were selected.

[12] Some of these dubious equivalents are actual translations in the original corpus, such as the Spanish-English *copa de champaña* 'champagne cup', which was translated as *cup of wine*, even if they are semantically different.

Table 2: Number of bilingual extractions of the baseline, NAT, and DIS systems.

Pattern	model	ES-PT	ES-EN	PT-EN
AMOD	BAS	657	248	213
	NAT	1,329	1,113	1,005
	DIS	9,464	7,778	7,083
NMOD	BAS	320	32	43
	NAT	704	138	136
	DIS	3,867	890	917
VOBJ	BAS	529	183	241
	NAT	1,443	1,461	1,544
	DIS	12,887	8,865	9,206
VSUBJ	BAS	188	27	55
	NAT	382	346	323
	DIS	2,522	1,344	1,298
ADVMOD	BAS	58	19	22
	NAT	113	104	106
	DIS	3,721	2,301	2,412

recall for the AMOD pattern was estimated as follows (see Table 1, Table 2, and Table 3): $r = \frac{p*e}{i} = \frac{99*657}{12{,}967} = 5.01$.[13] Finally, f-score values (the harmonic mean between precision and recall) were obtained for each case, and the macro-average results were calculated for each language, pattern, and approach.

Table 2 contains the number of bilingual collocation equivalents extracted by each method in the 15 settings from the input lists of monolingual data (Table 1). These results clearly show that the baseline approach extracts a lower number of bilingual equivalents. NAT obtains much more bilingual collocations than BAS, but both methods extract less equivalents than the distributional approach. This might have happened due to the size of the dictionaries in BAS and because of

[13]Note that these recall results assume that every collocation in the shortest input list of each pair has an equivalent on the other language, which is not always the case. Thus, more realistic recall values (which would need an evaluation of every extracted pair) will be higher than the ones obtained in these experiments.

Table 3: Precision, recall and f-score of the baseline (BAS) system (*average* is macro-average).

Pattern		ES-PT	ES-EN	PT-EN	average
AMOD	P	99.0	95.8	97.9	97.6
	R	5.0	1.7	1.6	2.8
	F1	9.6	3.4	3.2	5.4
NMOD	P	97.8	100	91.7	96.5
	R	5.5	1.0	1.3	2.6
	F1	10.5	2.0	2.5	5.1
VOBJ	P	98.7	100	92.1	96.9
	R	3.0	1.2	1.4	1.9
	F1	5.7	2.3	2.8	3.6
VSUBJ	P	93.8	96.3	92.7	94.3
	R	4.5	1.0	1.9	2.5
	F1	8.6	1.9	3.8	4.8
ADVMOD	P	96.7	100	95.7	97.4
	R	1.0	0.3	0.3	0.6
	F1	2.0	0.7	0.6	1.1
average	P	97.2	98.4	94.0	96.5
	R	3.8	1.0	1.3	2.1
	F1	7.3	2.1	2.6	4.0

the internal properties of the collocations in both BAS and NAT, where the collocates may not be direct translations of each other. Moreover, with all three strategies, the bilingual extractions including English are smaller than the Spanish-Portuguese ones.

Concerning the performance of the three approaches, Table 3 (BAS), Table 4 (NAT), and Table 5 (DIS) contain the precision, recall and f-score for each language pair and collocation pattern. BAS obtains high-precision results for every language and collocation pattern (91.7% in the worst scenario), with a macro-average value of 96.5%. These results are somehow expected due to the quality of the hand-crafted dictionaries. However, because of the poor recall numbers, the general performance of BAS is low, achieving F-scores around 4%. Interest-

Table 4: Precision, recall and f-score of the probabilistic (NAT) system (*average* is macro-average).

Pattern		ES-PT	ES-EN	PT-EN	average
AMOD	P	92.5	92.5	83.3	89.5
	R	9.5	7.4	6.5	7.8
	F1	17.2	13.8	12.0	14.3
NMOD	P	91.1	98.7	91.4	93.7
	R	11.4	4.4	4.0	6.6
	F1	20.2	8.3	7.6	12.1
VOBJ	P	95.2	80.0	92.7	89.3
	R	7.8	7.5	9.2	8.2
	F1	14.3	13.8	16.8	15.0
VSUBJ	P	82.4	78.6	79.2	80.0
	R	8.0	10.2	9.6	9.3
	F1	14.6	18.1	17.1	16.6
ADVMOD	P	59.2	78.8	83.3	73.8
	R	1.2	1.5	1.3	1.3
	F1	2.4	2.9	2.6	2.6
average	P	84.1	85.7	86.0	85.3
	R	7.6	6.2	6.1	6.6
	F1	13.8	11.4	11.2	12.1

ingly, the size of the dictionary does not seem crucial to the results of the baseline. In this respect, the Spanish-Portuguese results are much better, especially in terms of recall, than Spanish-English, whose dictionary is more than twice as large. Also, the Portuguese-English results are slightly better than the Spanish-Portuguese ones, the latter being obtained using a dictionary built by transitivity.

The use of probabilistic translation dictionaries (NAT) increases the recall by a factor of more than three when compared to the baseline, but with a cost in precision, which drops, in average, from 96.5% to 85.3%. However, these differences allow the NAT approach to obtain much better F-scores than BAS. When looking at the different collocation patterns, it is worth noting that while AMOD, NMOD, and VOBJ have precision values of about 90%, VSUBJ, and especially ADV-

Table 5: Precision, recall and f-score of DiS system (*average* is macro-average).

Pattern		ES-PT	ES-EN	PT-EN	average
AMOD	P	92.9	92.0	90.5	91.8
	R	67.8	51.6	49.5	56.3
	F1	78.4	64.3	64.0	68.9
NMOD	P	93.8	88.0	90.0	90.6
	R	64.3	25.0	26.3	38.5
	F1	76.3	38.9	40.1	51.9
VOBJ	P	90.1	84.0	83.9	86.2
	R	66.0	48.1	49.9	54.7
	F1	76.5	61.2	62.6	66.7
VSUBJ	P	80.3	81.2	74.1	78.5
	R	51.6	41.0	36.1	42.9
	F1	62.8	54.5	48.6	55.3
ADVMOD	P	77.6	83.3	67.4	76.1
	R	52.2	34.7	24.4	37.1
	F1	62.4	49.0	35.8	49.1
average	P	86.9	85.7	81.2	84.6
	R	60.4	40.1	37.3	45.9
	F1	71.3	53.6	50.2	58.4

MOD (also with very low recall values) do not surpass 80% (with one case, ES-PT, with < 60%). As it will be shown in §4.2, some preprocessing issues might be the source of the some errors of ADVMOD extractions.

As for the DiS model, its precision is again lower than BAS and very similar to the NAT approach, with average results of 84.6%. However, the distributional strategy finds much more bilingual equivalents than the dictionaries, so recall values increase to an average of more than 45%. Again, VSUBJ and ADVMOD show worse precision values than the other three patterns. Besides, the NMOD extractions of the pairs including English have very low recall when compared to the other results. This might be due to not extracting nouns analyzed as COMPOUND (§3.1). As for the other two methods, the DiS Spanish-Portuguese results are bet-

ter than the two other language pairs, so the linguistic distance seems to play an important role in bilingual collocation extraction.

The method proposed in this chapter assigns a confidence value (obtained from the cosine similarity between the vectors of the base and the collocate equivalents) to each bilingual pair of collocations. In this respect, Figure 1 plots the average performance and confidence curves versus the total number of extracted pairs. This figure shows that by using a high confidence value (> 90%), it is possible to extract about 40,000 bilingual pairs with a high degree of precision. Besides, filtering the extraction with confidence values higher than 90% does not increase the precision of the system. This suggests that the errors produced in the most confident pairs arise due to factors other than semantic similarity, such as different degrees of compositionality.

However, as the confidence value decreases, the precision of the extraction also gets worse, despite the rise in the number of extractions which involves higher recall and consequently better f-score.

Finally, all the bilingual collocations extracted by DiS were merged into a single list with the three languages, thus obtaining new bilingual equivalents (not extracted directly by the system) by transitivity.[14] This final multilingual resource has 74,942 entries, 38,629 of them with translations in all three languages.

4.2 Error analysis

The manually annotated lists of bilingual collocations were used to perform an error analysis of the DiS system. These errors were classified in five types depending on their origin. Table 6 contains, for each error type, the macro-average rates of each collocation pattern as well as the final distribution of the error typology.

1. **Bilingual model (*BiModel*)**: Though useful, the bilingual word embedding approach produces some errors such as the identification of antonyms that have a similar distribution, which can align opposite collocation equivalents, such as the Portuguese-English pair $tecido_{base}$ $vivo_{collocate}$ = $tissue_{base}$ $dead_{collocate}$, instead of *living tissue*, where the extracted equivalent of the collocate *vivo* ('living' – in this context – or 'alive', in Portuguese) was *dead*. In most cases, however, the system obtained similar (but not synonymous) collocations, such as $chá_{base}$ $preto_{collocate}$ 'black tea' in Portuguese aligned to $coffee_{base}$ $black_{collocate}$ 'black coffee' in English.

[14] The merging process obtained 6,969 new bilingual collocation equivalents not present in the original extractions, and it also includes more than one translation for some collocations.

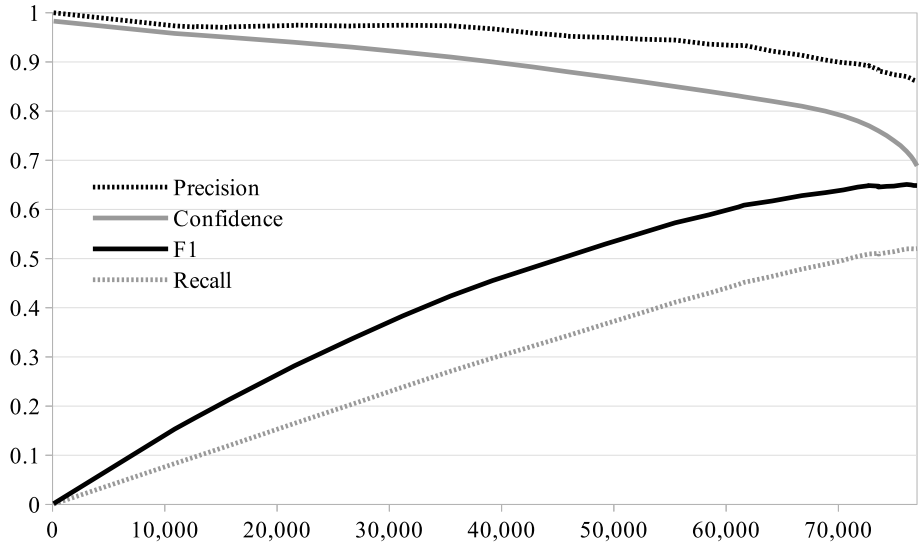

Figure 1: Average precision, recall, f-score, and confidence curves (from 0 to 1) versus total number of extractions of the DiS model.

2. **Monolingual extraction (*MonoExtract*)**: The extraction of base and collocate pairs produced incorrect collocations such as $plan_{base}$ $figure_{collocate}$, instead of obtaining the phrasal verb *figure out* as collocate.

3. **Preprocessing (*NLP*)**: Several errors derived from issues produced by the NLP pipeline, such as POS-tagging or dependency parsing: e.g., $pain_{\text{Noun}}$, end_{Verb} was labeled as DOBJ (instead of NSUBJ). A special case of preprocessing errors was the analysis of some Portuguese and Spanish adverbs ending in *–mente* (*-ly* adverbs in English), whose suffix was wrongly removed during the extraction process: e.g. *brutalmente* 'brutally' → *brutal*. These issues – which can be easily corrected – caused the alignment of incorrect Spanish and Portuguese collocations with English candidates, such as the Portuguese-English pair $matar_{base}$ $brutal_{collocate}$ = $kill_{base}$ $brutally_{collocate}$ instead of $matar_{base}$ $brutalmente_{collocate}$ = $kill_{base}$ $brutally_{collocate}$. This was the main source of errors of the ADVMOD relation.

4. **Lemmatization and gender (*Gender*)**: The lemmatization of some words differs from language to language, so working with lemmas instead of tokens also might involve some errors. For instance, the Spanish word *hija* 'daughter' is lemmatized as *hijo* 'son' (also in Portuguese: *filha*, *filho*),

while in English *son* and *daughter* appear as different entries. Thus, some bilingual collocations differ in the gender of their bases, such as the Spanish-English pair $hijo_{base}$ $encantador_{collocate}$ = $daughter_{base}$ $lovely_{collocate}$ instead of $hijo_{base}$ $encantador_{collocate}$=son_{base} $lovely_{collocate}$.

5. **Other errors (*Other*)**: Some other errors were caused by mixed languages in the original corpus. For example, the verb form *are*, in English, was analyzed as a form of the verb *arar* 'to plow' in Spanish. Some errors also arose from noise and misspellings in the corpora (proper nouns with lowercase letters, etc.).

It is worth mentioning that, in general, the error type distribution was similar across the different collocation patterns, showing much higher variation between different patterns of the same language pair. For instance, the distribution of Spanish-English AMOD errors is similar to the Portuguese-English AMOD one, while the typology of the Spanish-Portuguese NMOD errors is different to those of Spanish-Portuguese AMOD equivalents.

Table 6: Error rate of each of the defined error types of DiS system (*average* is macro-average).

Type	AMOD	NMOD	VOBJ	VSUBJ	ADVMOD	*average*
BiModel	70.57	93.52	59.23	45.74	32.61	60.33
MonoExtract	0	0	21.43	21.85	44.94	17.64
NLP	8.34	0	16.96	11.48	20.49	11.45
Gender	21.10	2.78	2.38	19.07	0	9.07
Other	0	3.70	0	1.85	1.96	1.50

Among the different errors produced by the presented method, an interesting case are *incongruent* collocations (Nesselhauf 2003). These expressions are those where the translation of both elements is not coherent, such as the English-Portuguese pair $requirement_{base}$ $meet_{collocate}$ = $condição_{base}$ $cumprir_{collocate}$, in which the verb *to meet* is usually translated into Portuguese as *conhecer*, not as *cumprir*. For these collocation equivalents to be correctly extracted by our method, they should appear with some frequency in the training corpus, which is not always the case. This fact may lead us to explore new compositional models, aimed at learning the distribution of the whole collocation, and not of its constituents, in further work.

4.3 Comparable corpora

A final experiment was carried out in order to find out (i) whether the bilingual word embeddings – trained on the same parallel corpora as those used for extracting the collocations – could be successfully applied to align collocations obtained from different resources, and (ii) the performance of the proposed method on comparable corpora.

Therefore, the same strategy for monolingual collocation extraction was applied in the Spanish and Portuguese *Wikipedia Comparable Corpus 2014*.[15] Then, we calculated the semantic similarity between the collocations using the same word embedding models as in the previous experiments.

From these corpora, filtered lists of 89, 285 and 140, 900 candidate collocations in Portuguese and Spanish were obtained, from 140M, and 80M of tokens respectively. From the 59, 507 bilingual collocations obtained by the DIS approach, 150 Spanish-Portuguese pairs were randomly selected and evaluated.

The precision of the extraction was 87.25%, with a recall of 58.15% (again computed using the whole set of monolingual collocations), and 69.79% f-score. These results are in line with those obtained on the OpenSubtitles Spanish-Portuguese pair (about 2% lower), so the method works well on different corpora and domains. It is worth noting that 49, 259 of the extracted collocation equivalents (83%) had not been retrieved from the OpenSubtitles corpus.

This last experiment shows that (i) the bilingual word embeddings can be used to identify collocation equivalents in different corpora than those used for training, and that (ii) they can also be applied to corpora of different domains to obtain previously unseen multilingual collocations.

5 Conclusions

This chapter presents a new strategy to automatically discover multilingual collocation equivalents from both parallel and comparable corpora. First, monolingual collocation candidates of five different patterns are extracted using syntactic analysis provided by harmonized UD annotation, together with a combination of standard association measures. Besides, bilingual word embeddings are trained on lemmatized parallel corpora. These bilingual models are then used to find distributional equivalents of both the base and the collocate of each source collocation in the target language.

The performed experiments, using noisy parallel corpora in three languages, showed that the proposed method achieves an average precision of about 85%,

[15] http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

with reasonable recall values. A systematic comparison to translation dictionaries pointed out that the distributional approach achieves similar precision results with much higher recall values than the probabilistic dictionaries. Furthermore, the evaluation showed that setting up a confidence value as a threshold is useful for retaining only high-quality bilingual equivalents, which could benefit the work on multilingual lexicography.

Finally, preliminary tests using comparable corpora suggested that the bilingual word embeddings can be efficiently applied to different corpora than those used for training, discovering new bilingual collocations not present in the original resources.

The multilingual resources generated by the proposed method can be used in several scenarios in which MWEs play an important role, such as machine translation or second language learning. In this respect, corpora from various registers and linguistic varieties could be used in order to obtain a wider diversity of collocation equivalents that can be useful for different purposes.

The work presented in this chapter enables us to propose a number of directions for further work. First, the results of the error analysis should be taken into account in order to reduce both the issues produced by the NLP pipeline, and those which arise from the word embedding models. On the one hand, understanding collocations as directional combinations may lead us to evaluate other association measures which are not symmetrical, e.g., *Delta-P*. On the other hand, it could be interesting to evaluate other approaches for the alignment of bilingual collocations which make use of better compositionality models, and which effectively learn the semantic distribution of collocations as single units, in order to deal with cases of incongruent collocation equivalents.

Abbreviations

EM	expectation maximization	NLP	natural language processing
EN	English	PT	Portuguese
MI	mutual information	ES	Spanish
MWE	multiword expression	UD	Universal Dependencies

Acknowledgements

This work has been supported by the Spanish Ministry of Economy, Industry and Competitiveness through the project with reference FFI2016-78299-P, by two *Juan de la Cierva* grants (FJCI-2014-22853 and IJCI-2016-29598), and by a 2017 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation.

References

Alonso-Ramos, Margarita, Leo Wanner, Orsolya Vincze, Gerard Casamayor del Bosque, Nancy Vázquez Veiga, Estela Mosqueira Suárez & Sabela Prieto González. 2010. Towards a motivated annotation schema of collocation errors in learner corpora. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the 7th Conference on International Language Resources and Evaluation* (LREC 2010), 3209–3214. European Language Resources Association (ELRA).

Altenberg, Bengt & Sylviane Granger. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics* 22(2). 173–195.

Berard, Alexandre, Christophe Servan, Olivier Pietquin & Laurent Besacier. 2016. MultiVec: A multilingual and multilevel representation learning toolkit for NLP. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC 2016), 4188–4192. European Language Resources Association (ELRA).

Church, Kenneth Ward & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22–29.

Evert, Stefan. 2008. Corpora and collocations. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. An international handbook*, vol. 2, 1212–1248. Berlin: Mouton de Gruyter.

Evert, Stefan & Hannah Kermes. 2003. Experiments on candidate data for collocation extraction. In *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics*, vol. 2 (EACL 2003), 83–86.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez & Francis M. Tyers. 2011. Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation* 25(2). 127–144.

Fung, Pascale. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas. Machine Translation and the Information Soup* (AMTA 1998), 1–17. Springer.

Garcia, Marcos & Pablo Gamallo. 2015. Yet another suite of multilingual NLP tools. In Sierra-Rodríguez, José-Luis and Leal, José Paulo and Simões, Alberto

(ed.), *Languages, applications and technologies. Communications in computer and information science* (International Symposium on Languages, Applications and Technologies (SLATE 2015)), 65–75. Springer.

Garcia, Marcos, Marcos García-Salido & Margarita Alonso-Ramos. 2017. Using bilingual word-embeddings for multilingual collocation extraction. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 21–30. Association for Computational Linguistics.

Garcia, Marcos, Marcos García-Salido & Margarita Alonso-Ramos. 2018. Discovering bilingual collocations in parallel corpora: A first attempt at using distributional semantics. In Irene Doval & María Teresa Sánchez Nieto (eds.), *Parallel corpora for contrastive and translation studies: New resources and applications*. John Benjamins Publishing Company.

Haruno, Masahiko, Satoru Ikehara & Takefumi Yamazaki. 1996. Learning bilingual collocations by word-level sorting. In *Proceedings of the 16th Conference on Computational Linguistics*, vol. 1 (COLING 1996), 525–530. Association for Computational Linguistics.

Hiemstra, Djoerd. 1998. Multilingual domain modeling in twenty-one. Automatic creation of a bi-directional translation lexicon from a parallel corpus. In *Computational linguistics in the Netherlands 1997: Selected papers from the eighth clin meeting*, 41–57.

Kilgarriff, Adam. 2006. Collocationality (and how to measure it). In Elisa Corino, Carla Marello & Cristina Onesti (eds.), *Proceedings of the 12th EURALEX international congress*, vol. 2, 997–1004.

Krenn, Brigitte & Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, 39–46. Association for Computational Linguistics.

Kupiec, Julian. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st annual meeting on association for computational linguistics* (ACL 1993), 17–22. Association for Computational Linguistics.

Lin, Dekang. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics* (ACL 1999), 317–324.

Lison, Pierre & Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Pro-*

ceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), 923–929. European Language Resources Association (ELRA).

Lü, Yajuan & Ming Zhou. 2004. Collocation translation acquisition using monolingual corpora. In *Proceedings of the 42nd annual meeting of the association for computational linguistics* (ACL 2004), 167–174. Association for Computational Linguistics.

Luong, Minh-Thang, Hieu Pham & Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the First Workshop on Vector Space Modeling for Natural Language Processing* (VSM-NLP) at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015), 151–159. Association for Computational Linguistics.

Mel'čuk, Igor A. 1998. Collocations and lexical functions. In Anthony Paul Cowie (ed.), *Phraseology. Theory, analysis and applications*, 23–53. Oxford: Clarendon Press.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013.* arXiv preprint arXiv:1301.3781.

Nesselhauf, Nadja. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied linguistics* 24(2). 223–242.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), 1659–1666. European Language Resources Association (ELRA). 23-28 May, 2016.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov & Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13(02). 95–135.

Orliac, Brigitte & Mike Dillinger. 2003. Collocation extraction for Machine Translation. In *Proceedings of ninth machine translation summit* (MT Summit IX), 292–298.

Pecina, Pavel. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation* 44(1-2). 137–158.

Rivera, Oscar Mendoza, Ruslan Mitkov & Gloria Corpas Pastor. 2013. A flexible framework for collocation retrieval and translation from parallel and comparable corpora. In *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology*, 18–25.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276/2010 (CICLing '02), 1–15. Springer-Verlag.

Seretan, Violeta. 2011. *Syntax-based collocation extraction* (Text, Speech and Language Technology). Dordrecht, Heidelberg, London, New York: Springer.

Seretan, Violeta & Eric Wehrli. 2007. Collocation translation based on sentence alignment and parsing. In *Actes de la 14e conference sur le traitement automatique des langues naturelles* (TALN 2007), 401–410. IRIT Press.

Simões, Alberto Manuel & José João Almeida. 2003. NATools – a statistical word aligner workbench. *Procesamiento del lenguaje natural* 31. 217–224.

Smadja, Frank. 1992. How to compile a bilingual collocational lexicon automatically. In *Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques*, 57–63. AAAI Press.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1). 143–177.

Smadja, Frank, Kathleen R. McKeown & Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics* 22(1). 1–38.

Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of language* 2(1). 23–55.

Wanner, Leo, Bernd Bohnet & Mark Giereth. 2006. Making sense of collocations. *Computer Speech & Language* 20(4). 609–624.

Wanner, Leo, Gabriela Ferraro & Pol Moreno. 2016. Towards distributional semant-ics-based classification of collocations for collocation dictionaries. *International Journal of Lexicography* 30. 167–186.

Wu, Chien-Cheng & Jason S. Chang. 2003. Bilingual collocation extraction based on syntactic and statistical analyses. In *Proceedings of the 15th Conference on Computational Linguistics and Speech Processing*, 1–20. Association for Computational Linguistics and Chinese Language Processing.